Journal of Experimental Psychology:
Learning, Memory, and Cognition
2004, Vol. 30, No. 3, 615–625

In the public domain
DOI: 10.1037/0278-7393.30.3.615

Testing Theories of Recognition Memory by Predicting Performance Across Paradigms

David G. Smith
Defence Research and Development Canada Toronto

Matthew J. J. Duncan
University of Western Australia

Signal-detection theory (SDT) accounts of recognition judgments depend on the assumption that recognition decisions result from a single familiarity-based process. However, fits of a hybrid SDT model, called dual-process theory (DPT), have provided evidence for the existence of a second, recollection-based process. In 2 experiments, the authors tested predictions of DPT and SDT by comparing the invariance of parameter estimates between yes/no (Y/N) and 2-alternative forced-choice (2AFC) testing paradigms. Both experiments showed DPT recollection estimates in Y/N to be poorly correlated with recollection estimates in 2AFC. In Experiment 2, SDT predictions explained more variance than DPT predictions. The authors evaluate and discuss the extent to which each model possesses theoretical validity versus computational flexibility in curve fitting.

Signal-detection theory (SDT) is one of the most common decision theories used in the analysis of recognition data (for a thorough introduction, see Green & Swets, 1966; Macmillan & Creelman, 1991).[1] The theory is used analytically to derive estimates of d' (sensitivity or strength) and bias (response criterion) from hit and false-alarm rates. Although there are problems and important considerations that must be addressed (see Lockhart & Murdock, 1970; Murdock, 1965), the theory can provide a competent description of the data. More recently, it has been proposed that retrieval in recognition is characterized by a second recollection process that operates in conjunction with a strength-based process (e.g., Mandler, 1980). From this dual-process account, a modified version of SDT has been introduced that combines a strength-based SDT-like process and a threshold-recollection process (e.g., Yonelinas, 1994). This new decision theory, known as Dual-Process Theory (DPT), has also been used analytically to derive estimates of d', response bias, and the probability of recollection (Ro) from hit and false-alarm rates (Yonelinas, 1994; Yonelinas, 1997; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996).

The successes of SDT and DPT in fitting recognition data have been used, respectively, as arguments for supporting either single-process or dual-process theories of recognition. Although several researchers have resorted to comparing the statistical quality of one model's fit with another's, the results of fitting the models are too ambiguous to interpret. For example, Yonelinas (1994; 1997; 1999) has found that DPT fits some data better than SDT, but others (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Slotnick, Klein, Dodson, & Shimamura, 2000) have found cases in which SDT seems a superior fit. Despite these demonstrations, there are at least two reasons why it is unlikely that arguments over subtleties of model fit can convincingly determine the superiority of one model over the other.

First, aside from model-specific theoretical properties that might improve characterization of the data, the fit of a model shows only that it is competent for a particular data set, not why the model fits (Myung & Pitt, 2002; Roberts & Pashler, 2000). For example, a model could fit merely because it is functionally more flexible than other models or because it contains more free parameters (see Myung & Pitt, 2002). Because model fit is a necessary but not sufficient condition, parameter values must have a sound theoretical interpretation if the model is also to be a psychologically plausible theory (i.e., the model must have face validity). However, even if these conditions are met, they do not imply that the model has good theoretical generalization beyond a particular data set. For this to be true as well, face validity must be extended to comparisons across experimental manipulations (for an extended review of validity issues, see Batchelder & Riefer, 1999).

Second, SDT and DPT are often applied in situations in which parameter values derived from fitting are not actually predicted, on a priori theoretical grounds, by the models themselves. The reason is that both SDT and DPT are typically used to represent the decision end of a more complex theory of memory from which testable predictions are actually formulated (e.g., SDT as the R System in the theory of distributed associative memory, Murdock, 1993). The situation is one in which manipulations across treatment conditions are assumed to affect memory but leave the

David G. Smith, Command Effectiveness and Behaviour Section, Defence Research and Development Canada Toronto; Matthew J. J. Duncan, School of Psychology, University of Western Australia, Crawley, Australia.

This research was supported by Grant 146 from the Natural Science and Engineering Research Council of Canada to Bennet Murdock. We thank Bennet Murdock for helpful comments on a draft of the article. This research was performed while both authors were at the University of Toronto.

Matthew J. J. Duncan is now at the Command Effectiveness and Behaviour Section, Defence Research and Development Canada Toronto.

Correspondence concerning this article should be addressed to David G. Smith, DRDC Toronto, 1133 Sheppard Avenue West, P.O. Box 2000, Toronto, Ontario, M3M 3B9 Canada. E-mail: david.smith@drdc-rddc.gc.ca

[1] Our generic reference to SDT applies to both the equal- and unequal-variance models. Where the distinction is crucial, we make it clear which model, equal- or unequal-variance, to which we are referring.

105-05773/

decision process unchanged (assuming, of course, that we compare the same decision task across conditions). The role of the decision theory is therefore that of an analytic tool used to generate parameter estimates. As a consequence, the face validity of those estimates must be assessed in terms of the memory theory from which they were derived and not in terms of the decision theory that generated the actual values. Testing SDT and DPT as psychological theories therefore requires experimental conditions in which predictions can be theoretically derived from the decision models themselves.

For the experiments reported here, we created such conditions by inverting the usual memory-decision relationship described above. By holding effects on the memory system constant while changing the decision rule across conditions, we allowed for parameter values to be assessed in terms of predictions made by the decision theory. As a test case for SDT and DPT, we compared performance between single-item yes/no (Y/N) and two-alternative forced-choice (2AFC) recognition tasks from a neutral encoding condition.

Obtaining reliable estimates for all the parameters in each model requires a minimum of two, but preferably six or more, hit and false-alarm rate pairs. Fitting either decision theory to multiple hit and false-alarm rate pairs amounts to fitting the theories to what is known as a receiver operating characteristic (ROC) curve (see Green & Swets, 1966; Macmillan & Creelman, 1991; Olgilvie & Creelman, 1968). One can obtain the empirical ROC curve in a single session by requiring subjects to simultaneously maintain various levels of confidence, each of which corresponds to a different criterion location. One can manipulate criterion location by asking subjects to rate the confidence of their response according to a scale of values that correspond to points along the strength axis.[2]

Y/N Models for SDT and DPT

A result consistently observed in Y/N ratings data is an asymmetrical ROC curve (e.g., Glanzer et al., 1999; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992). An example of a typical nonsymmetrical Y/N ROC curve is shown in Figure 1D. The lack of symmetry in the ROC curve can be readily explained by parameters in both SDT and DPT, but the theoretical basis for the explanation differs between models.

Signal-Detection Theory

Memory performance is defined as d', the standardized difference between the old and new distributions (see Figure 1A). The asymmetry observed in the ROC can be explained by SDT if it is assumed that the variance of the old-item distribution is greater than the variance of the new-item distribution (see Figure 1C). The greater the skew, the greater the variance of the old distribution relative to the new distribution. Although the asymmetry of the observed ROC can complicate matters when one is inferring d' if an equal-variance SDT model is assumed, the result itself poses no problem in estimating or interpreting the parameters in the unequal-variance model. The value of d' is still a valid point estimate of the standardized difference between the old and new distributions.

Dual-Process Theory

In this account, subjects are assumed to base their responses on either a familiarity process or a recollective process. On all trials, the relative strength of items is assessed in the same way assumed by an equal-variance SDT model. However, there is also some probability that the study event will be recollected in a process assumed to be an accurate episodic retrieval of the item from memory. If a study item is recollected, the strength of the item is ignored, and subjects are assumed to make a high-confidence *old* response. In the most common form of the two-process account, it is assumed recollection is independent of the strength-based process, implying that a recollected item could be either weak or strong in the strength domain.

In Figures 1E and 1F, we show the distribution and ROC assumptions of DPT. The new-item distribution in Figure 1E is identical to the new-item distribution for SDT. The dashed line represents the strength distribution for all old items. Because some proportion of old items are recollected with probability *Ro*, the solid line represents the remainder of the distribution for the items that are not recollected (hence the distribution has a lower density). The effect of recollecting some old items is to change the starting point of the ROC curve. That is, the hit rate at the highest possible confidence is equal to *Ro*, and the rest of the curve is drawn out with the remaining portion of the old-item distribution (plotted against the full new-item distribution). As skew increases, the model adopts a larger value for *Ro*. This starts the ROC at a higher intercept, compensating for the increased skew (see Figure 1F). A comparison of Figure 1D and 1F shows that the ROC curves predicted by SDT and DPT are qualitatively very similar.

2AFC Models for SDT and DPT

In 2AFC, subjects are presented with one old and one new item on every trial, with the task of choosing the old item. Presentation of two items makes the task different from Y/N because now the judgment is relative. Consequently, the shape of the 2AFC ROC predicted by both models is considerably different from ROC curves typically observed with Y/N.

Signal-Detection Theory

We assume that decisions are based on the difference in strengths between each item (difference rule, see Macmillan & Creelman, 1991). The assumptions of a difference rule have been empirically supported (e.g., Duncan, 1999) and produce estimates that are analytically and empirically similar to an independent-observation rule (e.g., McElree & Dosher, 1989).

According to a difference rule, the underlying distributions for each response option (left or right) are now strength differences. As such, the variance for each distribution will be defined by the statistical property of summation for random variables; namely, that the variance of the sum (or difference, in this case) is equal to the sum of the variances. Hence, the variance of strength differ-

[2] An alternate procedure is to have subjects maintain a single criterion for a block of trials but change criterion between blocks. Although procedural differences are relevant for comparisons between threshold and continuous state models (e.g., Malmberg, 2002), they do not affect a comparison between SDT and DPT.

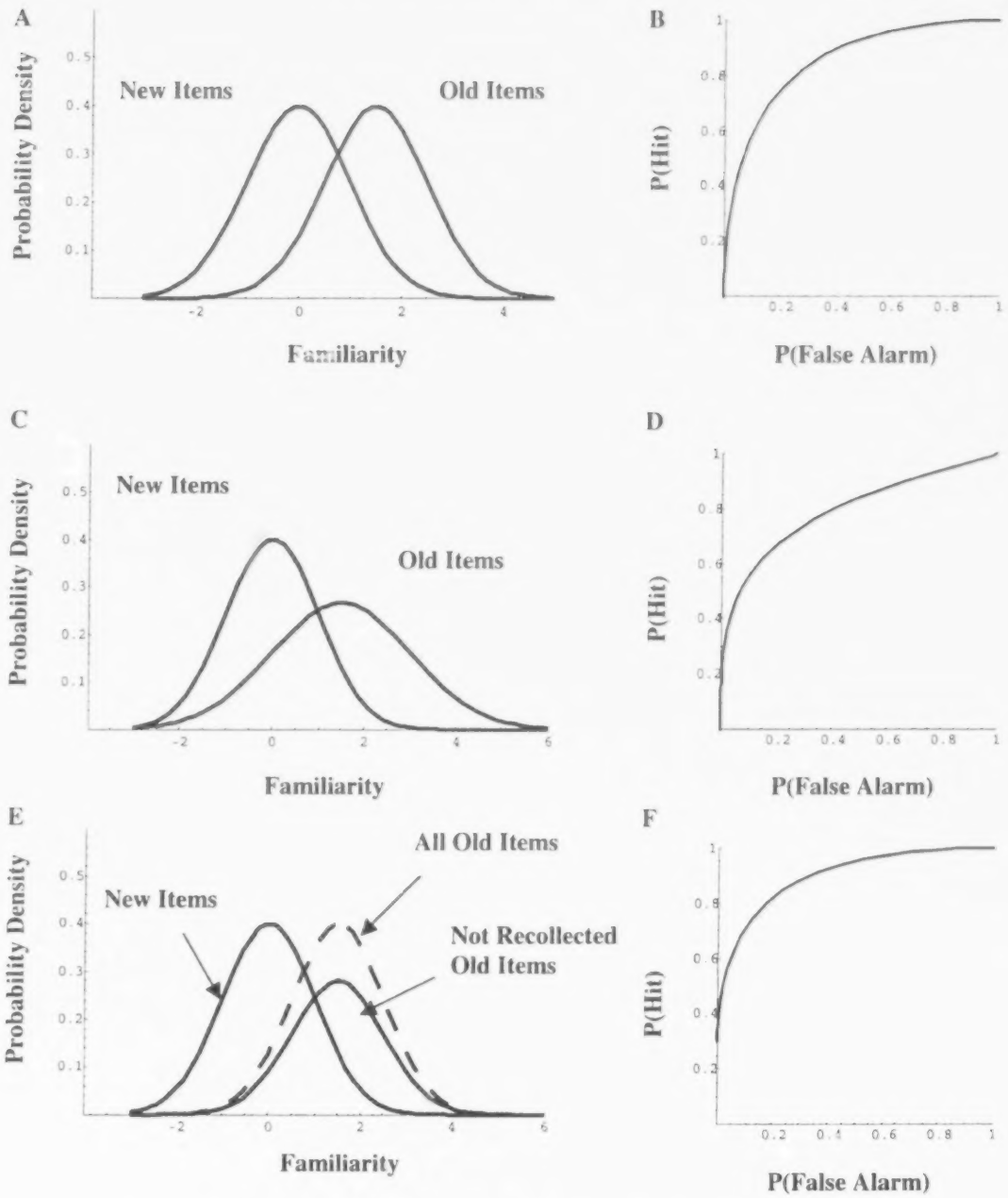

Figure 1. Strength distributions and receiver operating characteristic curves for equal variance signal-detection theory (SDT) (A, B), unequal variance SDT (C, D), and dual-process theory (E, F). FA = false alarm; P = probability.

ences will be equal for both response distributions (cases in which the old item is on the left or the right). Note that this is true regardless of whether the variances of individual old or new item-strength distributions are equal.

Unlike the Y/N model, the model for 2AFC makes a strong prediction about the shape of the ROC curve. Because the response distributions have equal variance, the ROC curve is predicted to be perfectly symmetric, like that shown in Figure 1B. Computation-

ally, one can accommodate observed deviations from symmetry by allowing the variance estimates for each distribution to be unequal (as in the Y/N model). But such an account is theoretically inconsistent with the model, despite the functional flexibility of SDT.

Dual-Process Theory

According to DPT, there are two possible events in a given trial. If the old item is recollected, the subject must respond confidently that the old item is old (recall that new items cannot be recollected). If the old item is not recollected, the subject resorts to an equal-variance version of the SDT difference rule. Consequently, if recollection were to fail on every trial, the ROC would correspond to a symmetrical curve like that in Figure 1B. However, if there is a probability higher than zero for recollection, then the start and endpoints of the ROC curve will divert from 0 on the *y*-axis and 1 on the *x*-axis, representing recollection probabilities for left or right items (see Equation 1 in Kroll, Yonelinas, Dobbins, and Frederick, 2002; see also Yonelinas, 1997; 1999).

Because of the familiarity process described by the equal-variance difference rule, deviations in symmetry will be reflected as a difference in left-right recollection probabilities. Like SDT, however, differences in left-right recollection are not theoretically predicted (see below) and so represent a functional, not a theoretical, competence.

Predictions Between Tasks

If decision processes are well described by SDT and DPT models, then performance in one task should be predicted from performance in the other. Specifically, formal properties of each model make it possible to define exactly what form the functional mapping of performance must take. Furthermore, neutral encoding of list items in each condition ensures that identical information is available for retrieval in either task. The combination of these two factors places a unique constraint on how parameter estimates in SDT and DPT are predicted to change across tasks.

Signal-Detection Theory

For SDT, the single source of information is assumed to be the familiarity or strength of the item (as estimated by d'). Because decisions in both tasks rely on the same kind of information, performance in one task should reflect an identical level of performance in the other. Given the two SDT models, d' in 2AFC is predicted from d' in Y/N by

$$d'_{2AFC} = 2\frac{d'_{YN}}{\sqrt{s^2 + 1}} \quad (1)$$

where s^2 is the variance estimate of the old-item distribution in Y/N (see Macmillan & Creelman, 1991, pp. 128). Essentially, this corresponds to a correction on Y/N performance to account for the different decision rule used in 2AFC.

Dual-Process Theory

The theory assumes two independent retrieval processes, so predictions are required for each. Familiarity-based retrieval, also parameterized as d', is equivalent to equal-variance versions of SDT for each task. We can therefore derive the same relationship for d' as expressed in Equation 1, however, because of the equal variance assumption $s^2 = 1$. As with SDT, the equation falls directly out of model assumptions and likewise represents a strong prediction for DPT.

For comparisons of *Ro*, the model makes no quantitative prediction under neutral encoding except for the identity function (e.g., that recollection in 2AFC equals recollection in Y/N). This prediction follows from the model because new items cannot be recollected. With a single recollectable item on each trial, the 2AFC decision rule for recollection remains unchanged, implying the identity function. Although the decision rule for familiarity is different under 2AFC, the two processes are assumed to be independent, and so no interference is predicted between them.

Kroll et al. (2002) have noted that estimates of *Ro* in 2AFC could be lower. The reason is that presentation of two items in the 2AFC task invokes emphasis on comparisons. As a consequence, subjects may opt to deemphasize recollected information; this results in lower recollection rates. A corollary to this is to assume that a 2AFC task leads subjects to base their decisions on a qualitatively different source of information.

Allowing strategic control over recollection as a direct attenuation of the recollection-process probability would mean that numerical estimates of *Ro* are actually measures of an arbitrary conscious strategy. The recollection parameter would become something akin to a strategic free parameter, not the measure of a lawful recollection process. If recollected information can be arbitrarily ignored, then contributions of recollection and familiarity to each response cannot be easily disentangled by the model. Hence, estimates of d' and *Ro* become theoretically ambiguous. Although it is possible to explain why DPT might produce different levels of recollection and sensitivity across different types of test, the explanation would weaken the model.

Previous Comparisons of SDT and DPT

The comparison of SDT and DPT across testing procedures was the basis of a series of studies by Kroll et al. (2002).[1] They showed that proportion correct (PC) in 2AFC could be predicted from DPT fits to Y/N data. However, there are limitations with the analysis conducted by Kroll et al. that we address in this article.

First, Kroll et al. (2002) did not report fits of DPT to the 2AFC data of Experiments 2A and 2B and so did not properly assess the face validity of either model's parameters directly. Second, and more important, they concluded that d' was an unreliable estimate of performance and opted instead to compare the models using PC. Despite the lack of a proper comparison of parameter estimates, the decision to use PC contributed to a serious confound to their analysis. Of grave concern is the fact that PC does not provide an unequivocal measure for interpreting the behavior of either SDT or DPT.

Green and Swets (1966) have shown that PC for unbiased observers in 2AFC is equal to the area under the Y/N ROC curve. However, both models fit the shape (and subsequent area) of the Y/N ROC using two main parameters (and an arbitrary number of criteria). With two parameters, there are many regular ROC curves, and hence many pairs of parameter values, that can give

[1] The studies by Kroll et al. (2002) were published after we had conducted our own experiments.

rise to the same area value (i.e., PC value). The main parameters of each model are therefore not identified with respect to PC. The weight of interpretation ultimately resides with model parameters. If an interpretation is to be unequivocal, model parameters need to be identified (Batchelder & Riefer, 1999; Myung & Pitt, 2002); this is a requirement clearly violated by using a congealed measure such as PC.

In addition to the interpretive problems of PC, the justification offered by Kroll et al. (2002) for choosing PC in the first place was unwarranted. Although it is true that SDT estimates of d' in Y/N do vary with response criterion, this is only true for an equal-variance SDT model. It poses no problems whatsoever for the unequal-variance SDT model assumed here. After decades of research producing skewed ROC curves in Y/N tasks, there is little doubt concerning which SDT model, equal- or unequal-variance, is the appropriate model for Y/N data.

We performed two experiments to test parameter invariance between tasks. We manipulated type of test (Y/N vs. 2AFC) between subjects in Experiment 1 and within subjects in Experiment 2. To maintain neutral encoding of the study list across test conditions, subjects in both experiments were not informed as to the nature of the tests until after the study phase. Experiment 1 served as a basic test of 2AFC model predictions as well as a point of reference for comparison with the within-subjects design used in Experiment 2.

Experiment 1

Method

Subjects. A total of 32 subjects participated in this experiment in return for course credit. Subjects were randomly assigned to either group so that each test condition contained 16 subjects.

Materials. A total of 280 words were randomly sampled from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982) and assigned as study items. For the Y/N test, 140 study items were randomly assigned as targets (old items), with 140 additional words sampled from the word pool to act as lures (new items). For the 2AFC test, all 280 study items were used as old items, with 280 additional words from the word pool assigned as new items. Each condition therefore consisted of 280 tests. Study and test items were shown in a black Times Roman 35-point font on a white background. Response keys used in the experiment were labeled on the keyboard.

Procedure. For the study phase, subjects were told that they would be asked to remember a word list consisting of 280 words presented on a computer screen. Each word was presented one at a time in the middle of the screen at a rate of 5 s per item. After viewing the list, test instructions were presented on the screen. For Y/N tests, subjects were told they would have to decide whether a single word was old or new and rate the confidence of their response. For 2AFC, subjects were told they would have to decide which of two words, one old and one new, was the old word and rate their response. Subjects were asked whether they understood the instructions before commencing with the test.

On each Y/N trial, a single word was presented in the center of the screen with the rating scale "(sure new) 3 2 1 1 2 3 (sure old)" underneath. Subjects were instructed to press "1," "2", or "3" on the right or left of the keyboard depending on whether they thought the word was old or new. Given their response (old or new), they were asked to choose "3" if they were sure, "2" if they were less sure, or "1" if they were unsure or guessing. For each 2AFC test, two words were presented in the center of the screen with the rating scale "(sure left) 3 2 1 1 2 3 (sure right)" underneath. Subjects were asked to press "1", "2", or "3" on the left or right, depending on which side they thought contained the old word, and to rate their response along the same scale described above. All subjects were told that test items would correspond to either response alternative with equal probability on any given trial. They were also asked to try to use each response key an equal number of times to scale their feelings of confidence[4]. Test order of the alternatives in each condition was randomized for each subject.

Results

We computed both group fitted and mean fitted parameter estimates. Mean fitted estimates were obtained by fitting individual subjects' ROC data and then averaging the parameter estimates. Group fitted estimates and the assessment of overall model fit were computed from a single group-averaged ROC curve. All ROC curves were fit using a search algorithm that maximized likelihood (see Olgilvie & Creelman, 1968).

SDT ROC analysis. Both SDT models were fit with 5 criteria, d', and the variance of the old-response distribution as free parameters (variance of the new-response distribution was set to 1.0). Figures 2A and 2B show the group fitted empirical ROC and group fitted parameter estimates for Y/N and 2AFC, respectively. Despite a good visual approximation to the data, both SDT models showed significant statistical deviation from the data, $\chi^2(3) = 16.14$, $p < .05$ and $\chi^2(3) = 17.85$, $p < .05$, for Y/N and 2AFC, respectively.

In Table 1, we show the mean fitted and mean predicted parameter estimates derived from each SDT model across the two tests. For d', the mean predicted value was computed from the average of individual subjects' fits. Despite marked differences, 2AFC d' predicted from Y/N was not significantly different from mean fitted 2AFC d', $t(27) = 0.900$, $p > .30$. Similarly, the mean fitted 2AFC variance was not significantly different from the theoretically predicted value of 1.0, $t(14) = 0.359$, $p > .50$. Finally, we used the group fitted Y/N estimate to compute a predicted value of 2AFC d'. The group predicted value of 2.06 was close to the actual group fitted value of 1.78.

DPT ROC analysis. The Y/N model was fit with 5 criteria, d', and $Ro_{Y/N}$, the probability of recollecting an old item. The 2AFC model required eight parameters: 5 criteria, d', Ro_L, and Ro_R. The latter two parameters represent the probability of recollecting an old item on the left or right. In Figure 2C, we show the fit of DPT to the grouped empirical ROC from the Y/N task. In Figure 2D, we show DPT model fit to the grouped empirical ROC from 2AFC. Grouped parameter estimates for each model are given in the figures. As with the SDT analysis, the data appear well fit by DPT. The Y/N model, $\chi^2(3) = 14.96$, $p < .05$, but not the 2AFC model, $\chi^2(2) = 2.44$, $p > .05$, showed significant deviations from the data.

Mean fitted estimates and the mean predicted values for each DPT model are given in Table 1. For d', the results of a t test between the mean estimated and mean predicted 2AFC values was not significant, $t(27) = 0.160$, $p > .80$. Mean estimated $Ro_{Y/N}$ was not significantly different from the average of mean estimated Ro_L (0.22), Ro_L (0.19), $t(27) = 1.608$, $p > .10$. Mean Ro_L and Ro_L

[4] Asking subjects to use each confidence criterion an equal number of times might seem contrary to DPT theory because recollected items would seem to demand a high-confidence response. However, subjects should be able to compensate as long as there is a large enough number of familiarity-based responses. Also, compliance to this instruction was not monitored during this experiment.

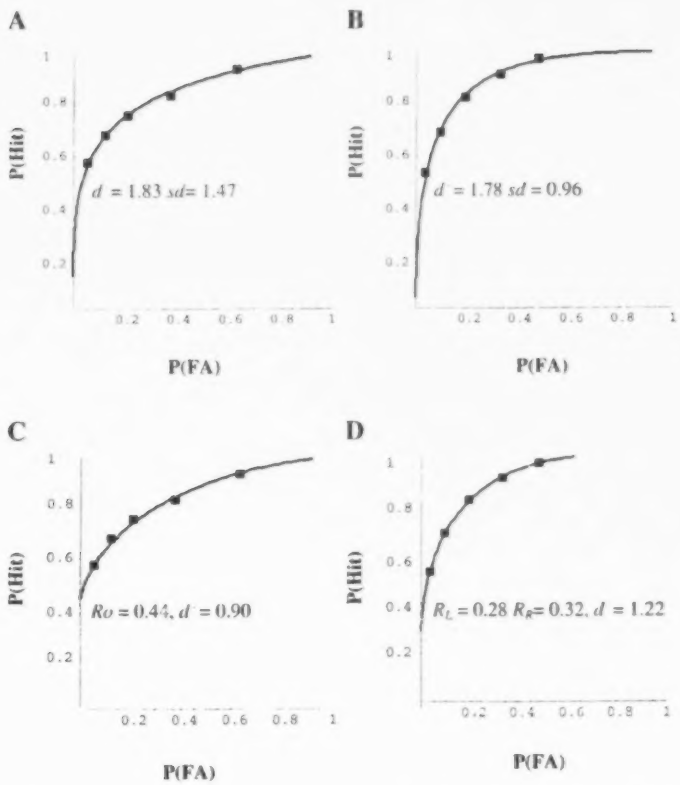

Figure 2. Empirical receiver operating characteristic curves from Experiment 1 for yes/no (left panels) and two-alternative forced-choice (right panels) fit with signal-detection theory (A, B), and dual-process theory (C, D). d' = standardized difference between stimulus class distributions (old vs. new for the Y/N task or old-right vs. old-left for the 2AFC task). FA = false alarm; P = probability; Ro = probability of recollection; R_L = probability of recollecting an item on the left; R_R = probability of recollecting an item on the right.

were not significantly different from each other, confirming the symmetry prediction.

Discussion

Both SDT and DPT captured the basic shape of the ROC curve in each condition. Although the model fits showed some statistical deviation, the probability of rejection with chi-square does increase with sample size. Considering the statistical power of Experiment 1, this is perhaps not too surprising. Because both models fit equally well, it also poses no problem for model comparisons. Results in the Y/N condition showed the typical finding of an asymmetric ROC curve. Both models fit the asymmetry in the usual way, with either a larger variance for the old-item distribution (SDT) or a nonzero recollection probability (DPT). In contrast, the ROC curve in 2AFC was symmetric. Symmetry was reflected as statistically equal variances for each response distribution (SDT), or as statistically equal left-right recollection probabilities (DPT). These results confirm the symmetry predictions and hence the validity of the 2AFC decision rules assumed by each model. When taken together, the models clearly fit each condition with theoretically consistent parameter values, especially in the case of 2AFC, and so meet the necessary conditions for addressing the pertinent question: namely, whether they produce parameter estimates that are invariant across paradigms. Both models answered this question less convincingly.

There was only weak correspondence of d' between the value estimated in 2AFC and the value predicted from the Y/N task. Although not statistically reliable, the predicted value was either higher (SDT) or lower (DPT) than the observed value. Recollection probability also did not show significant differences between the two tasks despite a large difference between predicted and observed parameters. Overall, the results suggest that model assumptions are incorrect or incomplete. However, there are other possibilities that explain the observed discrepancies.

One reason for the poor correspondence is that estimates were derived from subjects under different experimental procedures. For example, the 2AFC procedure contained twice as many individual words because each test trial contained a pair of test items. The excess items may have led to greater retroactive interference in the 2AFC task, which would have reduced performance resulting in

Table 1
Estimated and Predicted Model Parameters From Experiment 1

	Estimated		
Model	Y/N	2AFC	Predicted
SDT			
d'	2.11	1.87	2.22
sd	1.52	1.03	1.0
DPT			
d'	1.20	1.63	1.69
Ro	.35	.21[a]	.35

Note. Data from 1 subject in the two-alternative forced-choice (2AFC) condition could not be fit with the signal-detection theory (SDT) model, and data from 1 subject in the yes/no (Y/N) condition could not be fit with the dual-process theory (DPT) model. d' = standardized difference between stimulus class distributions (old vs. new for the Y/N task or old-right vs. old-left for the 2AFC task); Ro = probability of recollection.
[a] Represents the average of the mean fitted Ro_L and Ro_R estimates.

mispredictions from Y/N. In addition, test instructions for 2AFC often required more clarification than those for Y/N, causing slightly longer delays between study and test. The increased retention interval could have contributed to a further decrease in 2AFC performance, compounding the mispredictions found with the models. These possible confounds are especially relevant for explaining differences in *Ro* between tasks.

Experiment 2

The results of Experiment 1 demonstrated that SDT and DPT could readily fit data from each task, although both models were less successful in predicting performance across tasks. Hence, the goal of Experiment 2 was to modify our methodology in an attempt to improve predictability across tasks. To that end, we replicated the procedure used in Experiment 1 but varied test condition within subjects. This allowed for estimates and subsequent predictions to be calculated within individual subjects.

Method

Subjects. Thirty students from the same pool used in Experiment 1 volunteered in return for course credit. Two subjects were dropped because they failed to use the confidence ratings according to instructions.

Materials. Materials were the same as in Experiment 1, with a study list of 280 items. For Y/N, 70 study items were assigned as old, with 70 additional words from the pool assigned as new. For 2AFC, 140 of the remaining 210 study items were assigned as old, with 140 additional words from the pool assigned as new.

Procedure. After the study list, all subjects were given instructions for both tasks. They were also informed that there were equal numbers of each test and that test order for Y/N and 2AFC was randomized. Presentation and testing proceeded as in Experiment 1.

Results

Analysis and model-fitting procedures were carried out as described in Experiment 1. With test condition as a within-subjects factor, parameter estimates and corresponding 2AFC predictions were obtained for each subject individually. Because model predictions are derived from linear mapping functions in all cases, we evaluated reliability of model predictions using regression analysis.

SDT ROC analysis. The grouped empirical ROCs and corresponding fits of the model are shown in Figures 3A and 3B for Y/N and 2AFC, respectively. Tests of misfit to the grouped data were nonsignificant for both Y/N, $\chi^2(3) = 2.17$, $p > .05$, and 2AFC, $\chi^2(3) = 5.69$, $p > .05$. Mean fitted parameter estimates for each test condition, as well as corresponding predictions for 2AFC, are shown in Table 2. A test of the mean estimated variance in 2AFC against the predicted value of 1.0 was nonsignificant, $t(27) = 0.985$. Figure 4A shows the 2AFC d' predicted from Y/N d' for each subject plotted against observed 2AFC d'. The regression analysis produced a significant linear relationship, $R^2(27) = .55$, $p < .05$. Although significant, one reason for discrepancies between actual and predicted d' is that the 2AFC model assumes that left-right distributions have equal variance (i.e., the ROC is symmetric). Because the 2AFC variance estimate was not reliably different from 1.0, $t(27) = 0.985$, $p > .30$, we refit the 2AFC data, this time forcing equal variance. The results showed an improved correspondence between d' predicted from Y/N and that observed in 2AFC, $R^2(27) = .66$, $p < .05$.

DPT ROC analysis. Fits to the grouped ROC data, as well as group estimates of d' and Ro, are shown in Figure 3C for Y/N and Figure 3D for 2AFC. Visually, both decision models fit very well but still showed significant deviations from the data, $\chi^2(3) = 12.63$, $p < .05$, for Y/N and $\chi^2(3) = 7.67$, $p < .05$, for 2AFC. Mean fitted parameter estimates and mean predicted values are given in Table 2. For DPT, we conducted regression analysis for predictions of both d' and Ro.

A plot of d' predicted from Y/N performance against d' estimated from 2AFC performance is shown in Figure 4B. The analysis showed a significant linear relationship, $R^2(27) = .32$, $p < .05$. Figure 4C shows the corresponding estimates of $Ro_{Y/N}$ plotted against the average of left-right recollection probabilities, $(Ro_L + Ro_R)/2$, estimated from the 2AFC condition. It is clear from the figure that estimates of $Ro_{Y/N}$ did not predict recollection in 2AFC performance; this observation was confirmed by the resulting nonsignificant correlation, $R^2(27) = .002$, $p > .05$. The poor correspondence of recollection estimates between Y/N and 2AFC tasks may have been affected by the allowance of two different estimates of recollection (Ro_L and Ro_R) for 2AFC. Like left-right variance in SDT, estimates of left-right Ro must differ when there is asymmetry in the ROC, and so any left-right bias may have affected the prediction between tasks. Statistically, there was no significant difference between Ro_L (0.19) and Ro_R(0.18), $t(27) = 0.300$, $p > .70$. To double-check that this was not an issue, we refit the 2AFC data with a single recollection parameter (i.e., assuming $Ro_L = Ro_R$). The correlation between Y/N Ro and 2AFC $Ro_{L=R}$ improved from the original fit, but did not reach significance, $R^2(24) = .03$, $p > .05$.

Predicting PC

Recall that Kroll et al. (2002) found that PC in 2AFC could be predicted from PC in Y/N using either SDT or DPT. We tested this relationship using the same methods they used (i.e., Kroll et al., 2002, Equation 1). Performance in Y/N was fit to SDT and DPT, and the parameters were entered into Kroll et al.'s Equation 1 (the equation describing the PC relation for both models). Both SDT and DPT were able to predict 2AFC performance given Y/N

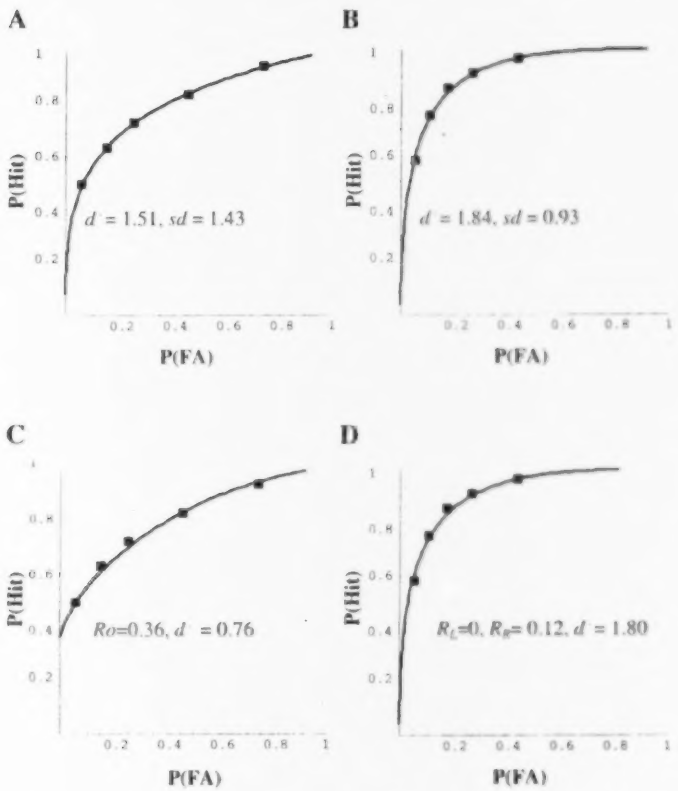

Figure 3. Empirical receiver operating characteristic curves from Experiment 2 for yes/no (left panels) and two-alternative forced-choice (right panels) fit with signal-detection theory (A, B), and dual-process theory (C, D). d' = standardized difference between stimulus class distributions (old vs. new for the Y/N task or old-right vs. old-left for the 2AFC task); FA = false alarm; P = probability; *Ro* = probability of recollection; R_L = probability of recollecting an item on the left; R_R = probability of recollecting an item on the right.

performance, $R^2(27) = .67$, $p < .05$, and $R^2(27) = .69$, $p < .05$, respectively.

Discussion

In this experiment, we used a within-subjects design in an attempt to improve parameter invariance between tasks and also to replicate the basic results found in the pure test conditions of Experiment 1. As in Experiment 1, we again found that ROC curves for Y/N were asymmetrical, whereas those for 2AFC were symmetrical (as predicted). Analytically, SDT showed an improved fit to the grouped data in each task separately, whereas DPT showed no qualitative improvements over Experiment 1. However, both models captured differences in shape of the ROC in ways consistent with those observed in Experiment 1.

The similarity of ROC curves between the two experiments is an indication that the presence of Y/N decisions did not have undue effects on 2AFC decisions (and vice versa). Although asymmetry in the 2AFC ROC increased in Experiment 2, the increase did not qualitatively change the predictive reliability of the models. Results of fitting the 2AFC ROC in Experiment 2 replicate those from Experiment 1 and lend further support to the decision rules assumed by the models for this task. We concluded that the decision rules remain unaffected whether testing occurs within subjects or within each task in isolation.

In Experiment 2, we adopted a within-subjects design to provide a more sound methodology for testing the invariance of parameter estimates between tasks. The analysis of SDT showed that this modification proved successful, as reliability of d' between tasks improved over that of Experiment 1. DPT also showed improved reliability in predicting d', but the predictions of SDT accounted for over twice the variance in observed 2AFC d' ($R^2 = .66$ vs. $R^2 = .31$, for SDT and DPT, respectively).

The gains in predictive consistency of d' made by DPT were subsequently offset by predictions of *Ro*. Adopting the improved methodology for parameter estimation actually produced worse predictive reliability of *Ro* compared with Experiment 1. Of special importance was the finding that an individual subject's retrieval via the recollection process in Y/N is uncorrelated with his

Table 2
Estimated and Predicted Model Parameters from Experiment 2

	Estimated		
Model	Y/N	2AFC	Predicted
SDT			
d'	1.65	2.13	1.82
sd	1.50	1.06	1.0
DPT			
d'	0.88	1.65	1.24
Ro	.29	.19[a]	.29

d' = standardized difference between stimulus class distributions (old vs. new for the yes/no (Y/N) task or old-right vs. old-left for the two-alternative forced-choice (2AFC) task). *Ro* = probability of recollection; SDT = signal-detection theory; DPT = dual-process theory.
[a] Represents the average of the mean fitted Ro_L and Ro_R estimates.

or her subsequent retrieval by recollection in 2AFC. According to Figure 4C, a group of subjects with observed recollection probabilities near .10 in 2AFC could have varied from near 0 to as high as .60 in their observed Y/N recollection probability.

The reliability differences between the models mentioned in the *Discussion* section were eliminated when predicting PC. In fact, both models showed the highest reliability of all in predicting this measure (DPT holding a marginal edge over SDT). A result such as this would appear to engender positive support for theoretical assumptions; this is an inference shared by Kroll et al. (2002). To the contrary, successful predictions of PC by both models actually demonstrate that PC provides no such unequivocal source of support.

General Discussion

In two experiments, we explored the theoretical relationship between Y/N and 2AFC recognition tasks within the predictions of SDT and DPT. Our rationale was to test SDT and DPT by obtaining empirical estimates of model parameters in 2AFC and then compare them with predictions theoretically derived from empirical Y/N estimates. We argued that theory-driven comparison of parameters was a more appropriate test than curve-fitting prowess or use of congealed measures such as PC. The comparison was used to determine which model produced more reliable predictions, and hence which was the more parsimonious psychological theory for these tasks.

In Experiment 1, we demonstrated that the form of the 2AFC ROC was consistent with that predicted by the decision rules of DPT and SDT. This result lends credibility to the predictions we derived for parameter estimates, because those predictions depend critically on the decision rules assumed by each model. Despite good fits to the separate data sets and subsequent nonsignificant tests of reliability, neither theory predicted parameter estimates between subjects that we considered theoretically consistent. In Experiment 2, we adopted a within-subjects design to improve precision. SDT was found to reliably predict *d'*, showing a significant correlation between tasks. DPT showed less, but still significant, reliability in predicting *d'*. Critically, however, levels of *Ro* estimated by the model in each task were uncorrelated with each other. But why then does the familiarity process in DPT not completely compensate?

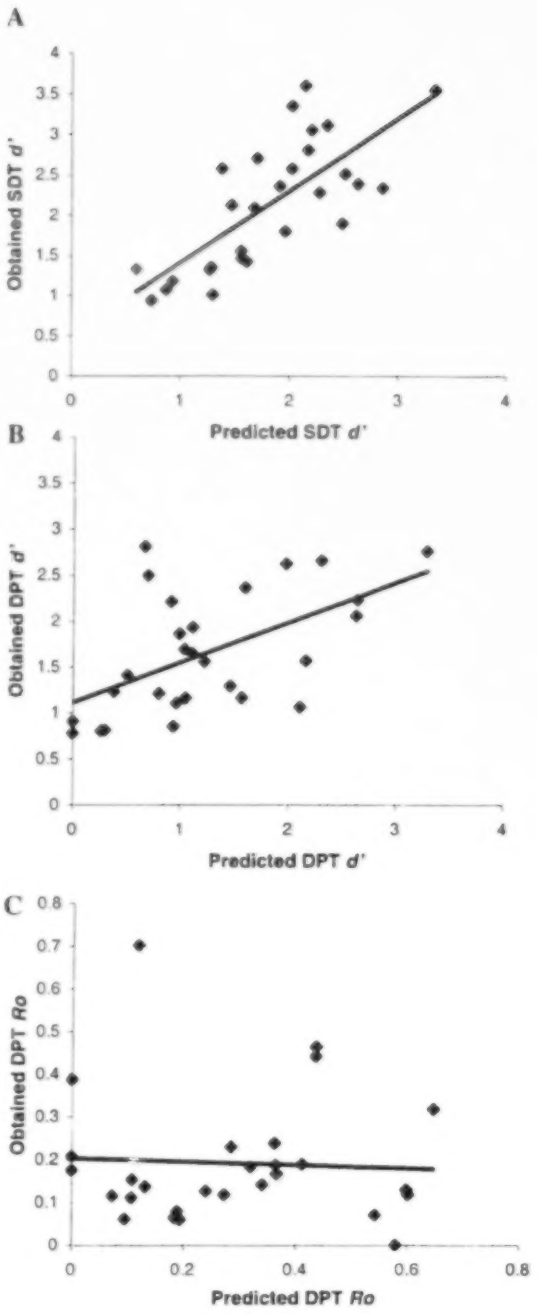

Figure 4. Scatter plots from Experiment 2. A shows predictions of two-alternative forced-choice (2AFC) *d'* from yes/no (Y/N) *d'* using SDT. B shows predictions of 2AFC *d'* from Y/N *d'* using DPT. C shows predictions of 2AFC *Ro* from Y/N *Ro* using DPT. *d'* = standardized difference between stimulus class distributions (old vs. new for the Y/N task or old-right vs. old-left for the 2AFC task). DPT = dual-process theory; SDT = signal-detection theory.

DPT is formally equivalent to SDT but for the recollection process and the equal variance assumption. Removing recollection from DPT would therefore make the models essentially equivalent (completely equal if we relax the equal variance assumption). Because recollection makes no theoretical contribution to predictability, the explanatory burden is placed squarely on the familiarity process (recall that the two are independent and so can make noninterfering contributions). The only theoretical difference between SDT and DPT (without recollection) is the equal-variance assumption of DPT. Although an equal-variance assumption will likely fit a 2AFC ROC reasonably well, it will clearly not do for the Y/N task. It therefore appears that the equal variance assumption is the main liability DPT has in predicting d'.

As we have seen, *Ro* makes no theoretical contribution to reliability across tasks, and so its contribution reflects the computational flexibility of DPT to fit ROC curves (as opposed to the explanatory contribution of d'). As we have argued, the fact that DPT has been used previously to describe the data (i.e., provide estimates) rather than to actually predict a priori the amount of recollection does not engender strong theoretical support for the model. In a decisive test relying on theoretical prediction, our own results and those of others (e.g., Buchner & Erdfelder, 1996) demonstrate that DPT assumptions are not sufficient for a complete picture of the role of recollective processes in recognition.

Conclusions

This research was motivated by the need to compare recognition-decision models as valid psychological theories, as opposed to their usual role as descriptive tools in the service of memory-level theories. In two experiments, we created experimental conditions in which explicit predictions could be derived straight from the theoretical assumptions of two highly successful decision theories: SDT and DPT. Although both models assume that decisions are based on item familiarity (parameterized as d'), DPT further assumes that an independent recollection process also contributes to recognition decisions (parameterized as *Ro*). The experiments involved comparing parameter estimates in one task (2AFC) with values theoretically predicted from parameter estimates in another task (Y/N). Our analysis showed that the predictions of SDT were the most reliably consistent with estimates derived from the data. In contrast, the predictions of DPT were either less reliable than SDT or uncorrelated with data-derived estimates.

Despite this, we do not consider the results of these experiments to be in any way decisive evidence against dual-process theories in general. We have constructed specific conditions tailored to permit the prediction and test of two models. Our results apply to the models tested, and we make no assertions regarding other decision theories. Importantly, this is not a limitation of testing parameter invariance across paradigms. Rather, it means that under conditions such as these, predictions are bound to be model-specific.

Given the specificity of our conclusions, we consider it likely that recognition decisions are based not only on multiple routes of retrieval (e.g., Brown, 1976) but on multiple sources of information as well (e.g., Banks, 2000). Clearly there is much work to be done, not only on theoretical grounds, but also in the procedures used to provide decisive theoretical support for models. We believe that the ability of models to predict across procedures is critical, and should, when possible, be emphasized over facility in curve fitting. With this in mind, although our present results support unequal-variance SDT, the theory may ultimately be shown as too simplistic in the long run.

References

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11,* 267–273.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Brown, J. (1976). *Recall and recognition.* Bath, England: Wiley.

Duncan, M. J. (1999). The dimensionality of memory strength between levels of both serial position and word frequency category. *International Journal of Psychology, 34,* 454–459.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation, 14,* 375–399.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500–513.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Huntington, NY: Robert E. Krieger.

Kroll, N. A., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131,* 241–254.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 380–387.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General, 118,* 346–373.

Murdock, B. B., Jr. (1965). Signal-detection theory and short-term memory. *Journal of Experimental Psychology, 70,* 443–447.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100,* 183–203.

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology Vol. 4: Methodology in experimental psychology* (pp. 429–460). New York: Wiley.

Olgilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

Ratcliff, R., Sheu, C.-f., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367.

Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 1499–1517.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25,* 747–763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434.

Yonelinas, A. P., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S., & King, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness and Cognition: An International Journal, 5,* 418–441.

Received April 2, 2002
Revision received September 8, 2003
Accepted November 18, 2003 ■